Spectra-based sentiment analysis

ALEXANDRE BARBOSA DE LIMA[1]
MARCOS R. PEREIRA BARRETO[2]
JOSÉ ROBERTO DE ALMEIDA AMAZONAS[3]

Faculty of Sciences and Technology, Pontifical Catholic University of São Paulo
Mechatronics and Mechanical Systems Department, School of Engineering, University of São Paulo
Telecommunications and Control Engineering Department, School of Engineering, University of São Paulo
Department of Computer Architecture, Universitat Politècnica de Catalunya
[1]ablima@pucsp.br
[2]mrpbarre@usp.br
[3]jra@lcs.poli.usp.br,amazonas@ac.upc.edu

Abstract. This paper introduces a method to identify the emotions conveyed by voice signals based on spectral analysis. The adopted tool is the periodogram associated to the measurement of the cumulative energy. It is shown that the variation of the emotional state produces a shift of the spectrum that can be taken as a emotional blueprint of the speaker and can be used to track his/her mood along the time.

Keywords: sentimental analysis, spectral analysis, periodogram, machine learning

(Received October 5th, 2018 / Accepted November 5st, 2018)

1 Introduction

1.1 Definition of sentiment analysis

Sentiment analysis demands a reference model for emotions. In other words: which are the emotions? Many models have been proposed; they can be divided basically into two categories: dimensional models and discrete models. Dimensional models search for the identification of the dimensions which defines emotions, as an analogy to position which can be defined in the Cartesian plane. Willhelm Max Wundt, recognized as the father of psychology, was the first to propose a 3-dimensions model in 1897 [4]. Most dimensional models are $2-$ or $3-$dimensional. Dimensional models with physiologic correlation have also been proposed, such as [5]. Discrete models, by they turn, search for the identification of emotions by enumeration. A fundamental debate, in this field, is the existence of an innate set of emotions, which would be cross-culturally recognizable. Paul Ekman model of the "Big Six" (happiness, sadness, fear, anger, disgust, surprise) is, perhaps, the most acknowledged result in this line [6]. The Ortony/CLore/Collins OCC Model is also a fundamental reference in this field, by defining twenty two (22) categories of emotions depending on the events, agents or objects in the environment [11].

1.2 Detection of emotion in voice signals as part of sentiment analysis

Emotion are conveyed in voice, gestures, semantic meaning and personal experiences. They are an integral part of human communication. Therefore, to build "intelligent computers" or even "sociable robots" [1], it is mandatory to give them the ability to recognize, understand and (perhaps) express emotions. This idea came from the seminal work of Rosalind Picard, defining the expression "affective computing" [14].

For this paper, only speech will be considered. The importance of emotional expression in speech communication has been recognized in History, with references in Aristotle and Cicero in their works in rhetorical [15]. Although multimodal emotion recognition [2] is an important subject, most research is unimodal, with facial

recognition being more frequent. But emotion recognition in speech is more and more relevant nowadays, with smart assistants such as Alexa and Siri being a trending topic.

1.3 State of art and trends in detection of emotion in voice signals

Most works on speech emotion recognition follow the same basic experimental procedure: a number of physical features is extracted from voice signal and an Artificial Intelligence algorithm is used to identify the conveyed emotion. The number of physical features varies from 100s to 1000s [17]. They include features directly extracted and transforms such as Mel Frequency Cepstral Coefficients (MFCC). Features are frequency- and intensity-related. The features are fed into a classifier, using algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), k Nearest Neighbors (kNN), Gaussian Mixture Model (GMM) and others [13], [9], [8], [18]. Each paper analyses different things, such as best algorithm, best feature set, intercultural differences, etc.

1.4 Contribution of the paper

This paper takes a step back into this process, to deeper analyse the dynamic characteristics of voice signals in presence of emotions. It has been motivated by searching for a justification for some well-known rules in Phonetics, such as those in [15], stating some acoustic patterns of basic emotions, such as intensity, mean f_0[1], f_0 variability, f_0 range, high frequency energy. For instance, it´s known that joy makes f_0 to increase whilst decreases in sadness. If a small set of feature characterizes emotion in speech, why do we have to use 100s or 1000s of features to identify emotions? Factorial experiments do not allow to reduce this quantity [17]. Therefore, a deeper reflection seems to be advisable.

The search is for an algorithm easy to implement, with a small set of features that can be computed in real time, considering the different timescales of speech, which are very distinct from facial or gesture. Perhaps, this search leads to a small set of features, amenable to fed into a machine learning algorithm.

1.5 Organization of the paper

This paper is organized in three sections, besides this. Section 2 discusses spectral analysis of random signals, with emphasis on periodogram and Daniell method, as periodograms were used as the main tool for sentiment analysis. Section 3 presents experiments conducted on the use of periodograms in sentiment analysis. Finally, Section 4 presents our conclusion and directions for future works.

2 Spectral Anaysis of Random Signals

2.1 The Notion of Time Series

Spectral analysis is part of time series analysis [12]. Therefore, we must first introduce the notion of time series.

Figure 1 presents and example of a time series. Roughly speaking, a time series consists in a set of numbers corresponding to the observation of a certain phenomenon. By nature, such numbers are random variables. This figure shows the speech signal corresponding to the vocalization of *aaa...hhh*. It is easy to observe a periodic behavior and a decrease of the power as the time increases [3].

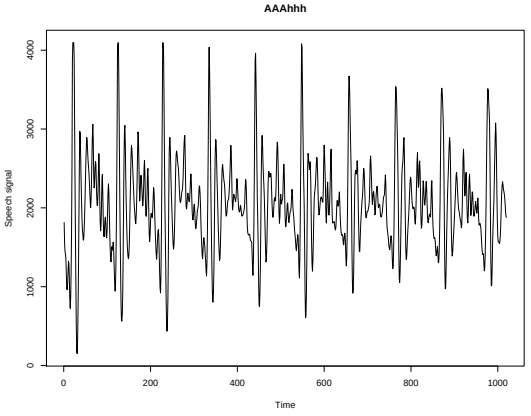

Figure 1: Speech signal: aaa...hhh. Source: [3]

2.2 Discrete Fourier Transform

Spectral analysis is a well-established research area. It continues to be a very useful tool in many areas of physical sciences such as oceanography, engineering, geophysics, astronomy and hydrology [12]. However, the

[1] A continuous time signal $x(t)$ with finite energy and duration T_0, where $\frac{1}{T_0} = f_0$ is the fundamental frequency (or fundamental harmonic) of $x(t)$, can be reconstructed using the Fourier Series [7]

$$x(t) = C_0 + \sum_{k=1}^{\infty} C_k \cos(k2\pi f_0 t + \theta_k)$$

$t_1 \leq t \leq t_1 + T_0$, $k = 0, \pm 1, \pm 2, \ldots$, where the amplitude C_k versus kf_0 defines an **amplitude spectrum**, the phase θ_k versus kf_0 is the **phase spectrum**, and the $|C_k|^2$ versus kf_0 is the **power spectrum**. These plots together are the frequency spectra of $x(t)$.

estimation of the power spectrum of a signal is not a trivial matter. There are two classes of spectral analysis techniques currently in use: parametric (or model-based) and nonparametric analysis.

The fundamental idea of parametric spectral analysis is fairly simple. The parametric approach assumes that the signal satisfies a generating model with known functional form and then proceed by estimating the parameters in the assumed model [16]. In this way, the parametric approach introduces the additional problem of model identification, and it is for this reason that we opt for the alternative technique, since the nonparametric approach (the one adopted in this article) is very useful for the purpose of this work (i. e., the search for an algorithm easy to implement, with a small set of features that can be computed in real time).

Let T be an arbitrary set. A stochastic process is a family $\{\boldsymbol{x}_t, t \in T\}$, such that, for each $t \in T$, $\boldsymbol{x}_t$ is a random variable [10]. When the set T is the set of integer numbers $\mathbb{Z}$, then $\{\boldsymbol{x}_t\}$ is a discrete time stochastic process (or random sequence); $\{\boldsymbol{x}_t\}$ is a continuous time stochastic process if T is taken as the set of real numbers $\mathbb{R}$. Then, the estimation of the Power Spectral Density (PSD) of a random sequence $\{\boldsymbol{x}_t\}$ can be made using periodogram methods based on the Discrete Fourier Transform (DFT), which can be efficiently calculated by an Fast Fourier Transform (FFT) Algorithm.

In the sequence we present the periodogram method used in this paper.

2.3 Power Spectral Density of Random Sequences

Consider a random sequence $\{\boldsymbol{x}_t\}$ with N values or samples, i. e., $\boldsymbol{x}_t = 0$ outside the time interval $0 \leq t \leq N-1$. In some cases of interest, we consider that $\{\boldsymbol{x}_t\}$ has a size N, even if its actual size is $M \leq N$ (in such cases the sequence $\{\boldsymbol{x}_t\}$ must be completed with $(N-M)$ zeros (zero padding).

The equation (1) is the DFT of $\{\boldsymbol{x}_t\}$:

$$X[k] = \begin{cases} \sum_{t=0}^{N-1} \boldsymbol{x}_t e^{-j2\pi \frac{k}{N} t}, & 0 \leq k \leq N-1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

2.4 The Periodogram

The PSD of $\{\boldsymbol{x}_t\}$ can be estimated by calculating the periodogram, given by

$$P_x(f_k) = \frac{|X[k]|^2}{N}$$

where $X[k]$ denotes the DFT of $\{\boldsymbol{x}_t\}$, and

$$f_k = 0, \left(\frac{1}{N}\right), \ldots, \left(\frac{k}{N}\right), \ldots, \left(\frac{N-1}{N}\right)$$

The periodogram is an asymptotically unbiased spectral estimator of the PSD of a random signal [16]. Its main problem lies in its large variance. In other words, the periodogram is inconsistent (i. e., the dispersion of the estimates is independent of N). This motivates the a "refined periodogram method", like the Daniell method.

2.5 Daniell Method

It is possible to show that the periodogram values $P_x(f_k)$ are asymptotically uncorrelated random variables [12]. Thus we may reduce its large variance by weight averaging the periodogram over small intervals centered on the current frequency f_k. The practical form of the Daniell estimate can be performed using the FFT. The Daniell kernel used in this worked is given in Section 3.2. For further details, please refer to [16], [12].

3 Results

In this section the simulation environment, the scenario setup and the simulations results are presented.

3.1 The simulation environment and scenario setup

All the results were obtained using the R-software that is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. It can be downloaded from `https://www.r-project.org`.

The simulation scenario is made of voice samples obtained from 22 different speakers, 50% women and 50% men. Each speaker produced 4 samples corresponding to different emotional states, namely: i) sad; ii) neutral; iii) happy and iv) angry. The emotional states were provoked by the texts offered to them to read. In this experiment, there was no means to check the real emotional state of the speaker and this may account for some discrepancies of the results, however this does not jeopardize neither the method nor the obtained results. All the voice samples were padded with zeros to have the same length of $n = 524,288 = 2^{19}$ points. For a sample rate of 8 kHz, this number of points corresponds to a 65.54 s long signal $s(t)$. All samples were normalized to have unitary energy, i.e., $\sum_{\mathrm{i}=1}^{\mathrm{n}} s(t)^2 = 1$.

3.2 Simulation results

The first step of the analysis consisted in obtaining an estimation of the signals' power spectral density. This was done by evaluating the periodogram using a Daniell's window with weights $\{h_k\} = \{\frac{1}{4}, \frac{2}{4}, \frac{1}{4}\}$ and a taper of 10%. Figure 2 shows a typical result. In this case the speaker is Allan and the emotion is *happy*.

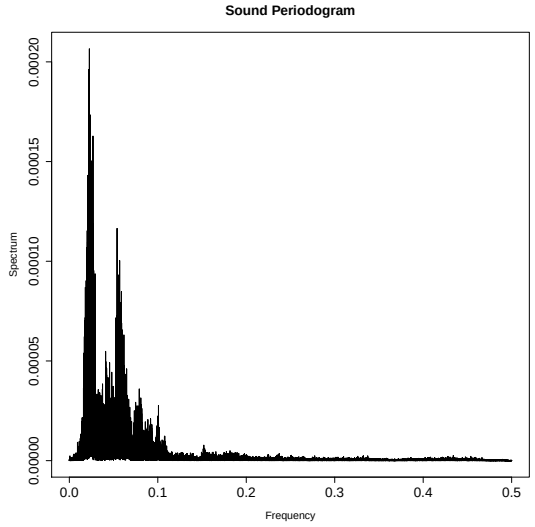

Figure 2: A typical periodogram: speaker - Allan; emotion - happy.

Figure 3 shows the periodograms for the four emotions of a typical speaker. Not surprisingly, it can be clearly seen that the power spectral density of the voice is affected by the emotional state. In fact, this result corroborates the common knowledge that when a person is *happy* or *angry* he or she tends to produce more higher frequency tones, while in the *sad* mood he or she tends to produce more lower frequency tones.

This result also suggests that the detection of the power spectral density shift may be a useful tool to identify the emotional state of a speaker.

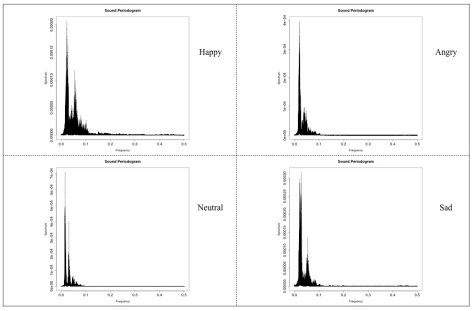

Figure 3: Periodograms for the four emotions of a typical speaker.

A convenient way to detect the power spectral density shift is by means of the cumulative energy given by $Cum\_En(k) = \sum_{i=1}^{k} PSD(k)$, where $PSD(k)$ is the power spectral density at the normalized frequency k. Figure 4 shows the cumulative energy for the four emotional states of a typical speaker. This figure clearly shows the spectral shift and each emotional state is characterized by a distinctive curve.

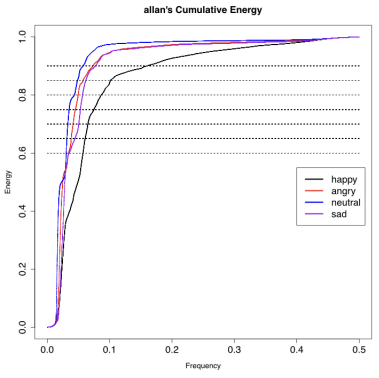

Figure 4: The cumulative energy of Allan's periodograms.

Figure 5 shows the cumulative energy of four speakers. In all cases we can realize that there is a spectral shift due to the variation of the emotional state. However, this shift is not exactly the same for all speakers.

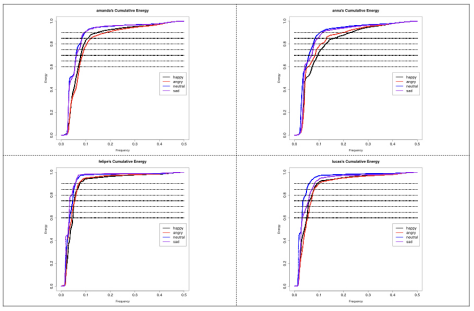

Figure 5: The cumulative energy of four speakers periodograms.

In Figures **??** and 5 the horizontal dotted black lines correspond to specific levels of the cumulative energy, namely, $Cum\_En(k) = \{0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90\}$. The measurement of the frequency at which each of these energy levels is reached constitutes the basis of the emotional state identification method outlined in Section 3.3.

3.3 Interpretation and discussion

Conceptually, the results shown in Figure 5 can be represented as illustrated in Figure 6. In this figure the horizontal axis represents $Cum\_En(k)$ and the vertical axis represents the normalized frequency. Each color represents an emotional state, and each *star* indicates the frequency at the specific cumulative energy level for the corresponding emotional state.

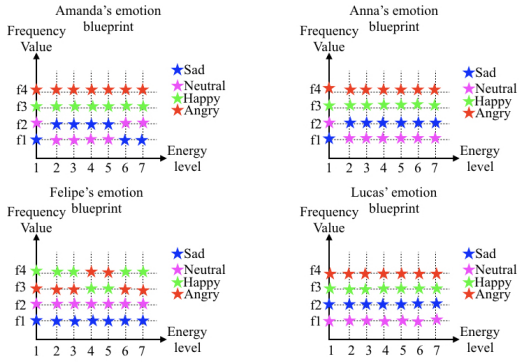

Figure 6: The emotional blueprints energy of four speakers.

This representation is designated as the *emotional blueprint of the speaker* because it can be assumed as a personal emotional identification of the speaker. For example, Lucas (bottom right figure) always remains in the same frequency curve whatever the cumulative energy is. When in the *neutral* mood the frequency curve is the lowest one and when in the *angry* mood the frequency curve is the highest. Amanda (top left), on the other hand, when in the *neutral* mood the frequency curve is the second lowest, as the energy level increase the frequency shifts to a lower level, but above $Cum\_En(k) = 0.6$ the frequency shifts back to the second lowest value.

These results suggest that spectra-based sentiment analysis can be implemented. Let's assume a stressful working environment as, for example, help desks or telemarketing, where the operators are under different stimuli for many hours. It may be important to detect the operator's emotional state to protect him/her from a psychological breakdown. Periodically, voice samples could be collected, the frequency reached for different cumulative energy levels evaluated and compared to the operator's blueprint enabling the identification of his/her emotional state. Alternatively, the measurement of the frequency could be made always at the same cumulative energy level and the observed shifts would indicate the variation of the emotional state.

4 Conclusions and Future Work

In this paper we have conceptually shown the possibility of identifying emotional states by means of the spectral analysis of voice signals. The outlined method, based on the cumulative energy of the power spectral density is quite parsimonious in terms of number of voice parameters to consider. We do not claim that the alternative presented in this paper is the most adequate as further and deeper research and comparisons are needed. However, it is clear that spectra-based analysis is a very promising path to be followed.

In future work we intend to:

- perform dynamic Fourier analysis and use wavelets transforms;
- introduce the measurement of energy as an additional tool to identify the emotional state;
- use the results of spectral analysis as input to machine and deep learning algorithms.

References

[1] Breazeal, C. Toward sociable robots. *Robotics and Autonomous Systems*, Vol. 42:167–175, 2003.

[2] Datcu, D. and Rothkrantz, L. Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition. *EUROMEDIA*, 2008.

[3] de Lima, A. B. and de Almeida Amazonas, J. R. *Internet Teletraffic Modeling and Estimation*. Gistrup: Rivers Publishers, 2013.

[4] Eerola, T. and Vuoskoski, J. K. The production and recognition of emotions in speech: features and algorithms. *Psychology of Music*, Vol. 39:18–41, 2011.

[5] Eerola, T. and Vuoskoski, J. K. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, Vol. 78:341–348, 2012.

[6] Ekman, P. An argument for basic emotions. *Cognition and Emotion*, Vol. 6:169–200, 1992.

[7] Fourier, J. Mémoire sur la propagation de la chaleur dans les corps solides. In *Présenté le 21 décembre 1807 à l'Institut national - Nouveau Bulletin des sciences par la Société philomatique de Paris, t. I, p. 112-116, n. 6*. march 1808.

[8] Kanchandani, K. B. and Hussain, M. A. Emotion recognition using multilayer perceptron and generalized feed forward neural network. *Journal of Scientific and Industrial Research*, Vol. 68:367–371, 2009.

[9] Kim, S., Georgiou, P. G., Lee, S., and Narayanan, S. Real time emotion detection system using speech: multimodal fusion of different timescale features. In *IEEE Workshop on Multimedia Signal Processing*. 2007.

[10] Morettin, P. A. *Econometria Financeira*. Associação Brasileira de Estatística, 2006.

[11] Ortony, A., Clore, G. L., and Collins, A. *The Cognitive Structure of Emotions*. Cambridge University Press, 1990.

[12] Percival, D. B. and Walden, A. T. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, 1993.

[13] Petrushin, V. A. Emotion recognition in speech signal: experimental study, development and application. In *Proceedings of the Internation Conference on Spoken Language Processing*. 2008.

[14] Picard, R. *Affective Computing*. MIT Press, 1997.

[15] Scherer, K. R. Vocal communication of emotion: a review of research paradigm. *Speech Communication*, Vol. 40:227–256, 2003.

[16] Stoica, P. and Moses, R. *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[17] Vogt, T. and Andre, E. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE Int. Conf. on Multimedia*. Netherlands, 2005.

[18] Vogt, T., Andre, E., and Bee, N. Emovoice-a framework for online recognition of emotions from voice. *Perception in Multimodal Dialogue Systems*, 2008.